DOCUMENT RESUME

ED 445 080 TM 031 757

AUTHOR White, Amy E.
TITLE Statistical Testing and Type I Error.
PUB DATE 2000-01-00
NOTE 16p.; Paper presented at the Annual Meeting of the Southwest Educational Research Association (Dallas, TX, January 27-29, 2000).
PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Hypothesis Testing; *Statistical Analysis
IDENTIFIERS *Type I Errors

ABSTRACT

In traditional null hypothesis testing, researchers use critical values of various test statistics in order to minimize the risk of making Type I errors. These critical values are associated with common alpha levels (e.g., 0.01, 0.05) that indicate the probability of a Type I error. Alpha values are set at conservative levels such that the Type I error is at a minimum for any given test. However, as the number of statistical tests conducted on a single sample increases, the change of making a Type I error somewhere in the study escalates appreciably, becoming much larger than the alpha level associated with any single test. An educational research data set is used to illustrate this problem, and several corrections are suggested, including better specification of research questions, increased parsimony in selection of statistical tests, and use of multivariate methods when possible. (Contains 3 tables and 16 references.) (Author/SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

Statistical Testing and Type I Error

ED 445 080

Statistical Testing and Type I Error

Amy E. White

University of North Texas

Amy E. White

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

TM031757

Paper presented at the annual meeting of the Southwest Educational Research Association, Dallas, TX, January 27-29, 2000.

BEST COPY AVAILABLE

Abstract

In traditional null hypothesis testing, researchers employ critical values of various test statistics in order to minimize the risk of making Type I errors. These critical values are associated with common alpha levels (e.g., .01, .05) that indicate the probability of a Type I error. Alpha values are set at conservative levels such that the Type I error is at minima for any given test. However, as the number of statistical tests conducted on a single sample increases, the chance of making a Type I error somewhere in the study escalates appreciably, becoming much larger than the alpha level associated with any single test. The author uses an educational research data set to illustrate this problem and suggests several corrections, including better specification of research questions, increased parsimony in selection of statistical tests, and utilization of multivariate methods when possible.

Statistical Testing and Type I Error: When Is Enough, Enough?

For over 70 years, researchers have employed analysis of variance in countless studies. The ANOVA, in use since 1925 (Fisher, 1925), has come to be one of the most widely used tests in social science research (Elmore & Woehlke, 1988). Most recently, statistical software like SPSS has made utilization of ANOVA and other parametric procedures even easier. However, technology often breeds dissent, and with the wide use of the ANOVA procedure disputes have come concerning the types of comparisons made with its use. For as useful as programs like SPSS are, they have also made the running of so-called post hoc tests easier, and in the end, overused. This form of unplanned comparisons, often referred to as "data snooping" (Kirk, 1969) or "milking" (Keppel, 1982) involves a host of post hoc tests. These tests are run after the initial ANOVA if the omnibus F is found to be statistically significant and consequently greatly increases the Type I error rate in these experiments. In essence, researchers with no real idea as to what they are searching for use post hoc tests to find some explanation for the statistical significance of the omnibus test, hoping to stumble upon some conclusion for which they can really take no credit having never supposed such a conclusion would occur. In addition, one of the many problems with this kind of research is that the experimentwise Type I error rate increases with every comparison. Would it not be better to know enough about one's population of interest to be able to plan one's comparisons before data are even collected? Some relevant literature would support just such a conclusion.

Review of Literature

While there is currently much debate about the uses and misuses of statistical significance tests in behavioral science research, (Daniel, 1998; Levin, 1998; Nix & Jackson, 1998; Thompson, 1998) these tests (SSΓ's) are still commonly used and must be understood and

used correctly to further the fields which are being researched. Thompson's 1990 paper thoughtfully illustrated the woes of planned or a priori comparisons versus post hoc testing when using analysis of variance (OVA) procedures in research. The problem, as Thompson points out, is that when researchers obtain a significant omnibus F, they must then locate the "cause" of the statistically significant result. After all, if only two means (K=2) are being evaluated, the task of determining where the difference lies is a much more simplistic task than when K>2 groups are beings analyzed. Thompson would argue that a thoughtful researcher carefully inspects the population of interest prior to conducting the study and plans the pairs to be tested by careful discernment. Other researchers have echoed this sentiment:

> The post hoc method is suited for trying out hunches gained during the data analysis. (Hays, 1981, p. 439)
>
> Post hoc comparisons are made in accordance with the serendipity principle—that is, after conducting your experiment you may find something interesting that you were not initially looking for. (McGuigan, 1983, p.151).

Other researchers would advocate the use of "planned comparisons" because these tests refine the research focus by statistically isolating the specific hypotheses to be tested (Schmitt, 1988). Paul Games' 1971 article carefully examines the effects of post hoc testing on the "familywise" (also called "experimentwise") Type I error rates. If a researcher has two independent tests, each run at an alpha of .05, the probability of rejecting both hypotheses (when they are true) is .95* .95 = $.95^2$. Because a rejection of either H_{o1} or H_{o2} causes a rejection of the overall null, the probability of a Type I error on the overall null is $1 - .95^2 = 1 - .9025 = .0975$ (Games, 1971). So by incurring two comparisons of means, the Type I error rate theoretically jumps from 5% to just over 9%.

Type I error rates are of great concern to this argument. As stated above by Games, the more comparisons that are run, the greater the chance of committing at least one Type I error in the research. Post hoc testing always increases the probability of occurrence of a Type I error. However, as Huck and Cormier (1996) point out, some of the most popular of these tests, Duncan's multiple range test; Fisher's LSD; the Newman-Keuls; the Scheffé and Tukey's HSD, all have some kind of built in correction for the Type I error rate. So, why does this not shed more positive light on the use of post hoc tests? The answer lies in statistical power.

The power or sensitivity of any given test is determined by a multitude of factors. For instance: the size of true treatment effects, the sample size, the degree of error variance and the selected statistical significance level all effect the power of a test (Benton, 1990). Keppel (1982) states the consideration of power should be among the first steps in planning any experiment. Because alpha levels and power are inversely related, when the aforementioned post hoc tests "adjust" for Type I errors, they are actually lessening the power of the test (the assurance of rejecting a null hypothesis that is, indeed, false). Keppel (1982) would go on to argue that in order to increase the power of a test, two common procedures are: (a) to increase the size of the sample, and (b) to employ an experimental design that provides a more precise estimate of treatment effects and a smaller error term. One way to achieve this better experimental design is the use of planned comparisons instead of post hoc tests. Tucker (1991) reiterated the thought this way:

> Because a given unplanned post hoc test corrects the alpha level for all the possible comparisons for a given study, even comparisons not of interest to the researcher or comparisons not even tested by the researcher, unplanned tests have less statistical power against

Type II error. (p. 111)

So, in searching the prevalent literature, it becomes evident that post hoc testing is less appropriate for ANOVA when K>2 in regards to both Type I and Type II error rates.

Practical Implications

Both Thompson (1990) and Benton (1990) examined the fallacious thinking and erroneous conclusions that can occur when omnibus F and post hoc tests are chosen instead of thoughtful, and therefore more meaningful, planned comparisons. In each instance, it becomes apparent that sometimes a statistically significant result on the omnibus F will yield no statistically significant groupwise comparisons in the post hoc tests. Likewise, these posteriori tests are only supposed to be used when an omnibus F actually is significant. The treatments of data by Thompson and Benton illustrate that it is indeed possible to have statistically significant differences in the means of two or more groups while still retaining the omnibus null hypothesis (a fact that would escape the attention of those choosing only ANOVA and post hoc comparisons). The following discussion with illustrations will serve to further these points and suggest ways to more effectively compare means when K>2 using a regression ANOVA and planned comparisons.

Planned Comparisons

As reviewed in the above literature, there are myriad reasons for a researcher to choose planned comparisons over post hoc testing. The four most obvious are:

a. planned comparisons lead to more thoughtful research
b. planned comparisons lower the risk of experimentwise Type I error
c. planned comparisons increase the power of the tests
d. planned comparisons help eliminate inconsistencies among tests (omnibus versus post hoc)

When using experimental research, it is assumed that the researcher wishes to gain as much credibility as possible to increase the confidence others will have in the research. One way to accomplish this is through the use of planned comparisons and a procedure known as "contrast coding", both of which will be discussed herein.

Data Examples 1

The ANOVA, along with follow-up post hoc tests, referenced in Table 1 was run using selected data from Daniel and King (1998, with permission). The original study, among other things, examined the effects of three inclusion groupings on various scores from standardized tests. When the three inclusion methods (EXPGROUP) were compared to the dependent variable SAT94TOT, some definite inconsistencies become evident.

For instance, in Table 1, the initial ANOVA shows the comparisons statistically significant at .05 level, but the effect size is rather small (10%), not a very appreciable or "important" result. However, when examining the post hoc tests, two of the three, 66%, of the comparisons show statistical significance at $\alpha = .05$, as the $P_{criticals}$ are <.001. Surely when so many of the tests show statistical significance, it would seem to indicate the likelihood that the effect size would be of greater magnitude, but clearly it is not. Secondly, although not a problem with this data set, there is frequently disagreement among the post hoc test results dependent upon which tests are selected. Not all post hoc tests will concur as to which results are statistically significant, and which are not. Inconsistencies in data interpretation is only one of the problems with comparisons that are not appropriately planned before the data are run.

Data Example 2

Another such problem, illustrated in Table 2, is the possibility of obtaining a statistically non-significant $P_{calculated}$ in the omnibus ANOVA, but when the post hoc tests are subsequently

run, one or more statistically significant results appear. Data used in this example are also from Daniel and King (1998, with permission); however, in this run, a random sample equal to 30% of the cases was examined. Though the omnibus test is not statistically significant ($p>.05$), one of the LSD follow-up tests is. (This example also illustrates the problem noted above regarding conflicting results dependent upon which post hoc test is employed.)

The danger here lies in the protocol for ANOVA routines. An omnibus F that is not statistically significant at the critical level for alpha, would theoretically be discarded. After all, post hoc tests should only be run after the omnibus null hypothesis has been rejected. Obviously, something has been missed in this scenario. This test's power against retaining a false null hypothesis has been so reduced that statistically significant pairwise comparisons have been overlooked. (Equally as heinous is the less occurring opposite where the F is statistically significant, put none of the post hoc tests is!) Such erroneous conclusions are of great danger to the researcher who does not plan comparisons before the data is run. Such a researcher is also increasing the theoretical Type I error probability. Remembering Games' calculations, we must account for all the comparisons actually being made in the post hoc tests:

»Group 1 & Group 2
»Group 1 & Group 3
»Group 2 & Group 3

Since now there are three comparisons and three null hypotheses, the formula becomes:

$$1 - .95^3 = 1 - .8573 = .1427$$

Though at some larger number of cases this formula ceases to be accurate, it is apparent from this result that the alpha level of .05 has now been expanded to slightly over 14%; a critical alpha from which many researchers and their readers would shy away. If knowing this, we then

attempt a so-called adjustment for the Type I error rate, we have accomplished only the decrease in the power of the tests. Seemingly this is an endless circle; or is it?

There are ways for thoughtful researchers to avoid such pitfalls while employing the use of multiple comparisons. One such method is the use of "contrast coding". In essence, the researcher codes variable values such that the groups which are deemed most likely to be the causation of the differences in means become the object of the statistical tests. Researchers should have some idea as to where the differences, if they exist, will occur among their groups of interest. Schmitt (1988) gives an excellent example of such a process. In the example shown in Table 3 (using hypothetical data simulating a comparison of exam scores from four experimental groups) data for a "balanced cell" ANOVA are coded to allow for three separate and planned contrasts. Each coding column uses a mean of zero and offsetting positive and negative values to indicate the group of groups to be compared. Data are then entered using a regression routine available in any given software package with each coding column entered at a subsequent block of the analysis. This procedure allows for the observation of results of the planned comparisons by comparing the differences in the regression sum of squares across subsequent blocks of the analysis.

Summary

Since ANOVA and other OVA procedures are in the statistical realm to stay, it behooves today's researcher to all but abandon post hoc testing in favor of planned comparisons. Not only will these planned comparison procedures help hold down the risk of Type I errors, but will also serve to increase the power and sensitivity of the tests. The use of planned comparisons also helps avoid inconsistencies in data analysis between the omnibus and post hoc tests. There is, however, an additional reason to employ a priori or planned tests; such procedures force

researchers to think. And there are few fields in existence today that could not benefit from more thoughtful treatments of research.

References

Benton, R. (1990). The statistical power of planned comparisons. Paper presented at the annual meeting of the Southwest Educational Research Association, Austin. (ERIC Reproduction Service No. ED314494)

Daniel, L.G. (1998). Statistical significance testing: a historical overview of misuse and misinterpretation with implications for the editorial policies of educational journals. Research in the Schools, 5, 23-32.

Daniel, L.G., & King, D.A. (1998). The impact of inclusion education on academic achievement, student behavior, student self-esteem, and parental attitudes: a multivariate investigation. Journal of Educational Research, 91, 331-344.

DuRapua, T. (1988). Benefits of using planned comparison rather than post hoc tests: a brief review with examples. Paper presented at the annual meeting of the Mid-South Educational Research Association, Louisville. (ERIC Reproduction Service No. ED303490)

Fisher, R.A. (1925). Statistical methods for research workers. Edinburgh, England: Oliver and Boyd.

Games, P.A. (1971). Multiple comparisons of means. American Educational Research Journal, 8, 531-565.

Huck, S.W., & Cormier, W.H. (1996). Reading Statistics and Research. NewYork: Harper Collins.

Keppel, G. (1982). Design and analysis: A researcher's handbook. Englewood Cliffs, NJ: Prentice-Hall.

Kirk, R.E. (1969). Experimental design: procedures for the behavioral sciences. Belmont, CA: Brooks/Cole.

Klockars, A.J., & Hancock, G.R. (1990). Competing strategies for planned orthogonal contrasts. (ERIC Reproduction Service No.ED319801)

Levin, J.R. (1998). What if there were no more bickering about statistical significance tests? Research in the Schools, 5, 43-54.

Nix, T.W., & Barnette, J.J. (1998). The data analysis dilemma: Ban or abandon. A review of null hypotheses significance testing. Research in the Schools, 5, 3-14.

Schmitt, D.R. (1988). The use of a prior techniques with a MANOVA. Paper presented at the annual meeting of the Mid-South Educational Research Association, Louisville. (ERIC Reproduction Service No.ED303504)

Thomson, B. (1998). Statistical significance and effect size reporting: Portrait of a possible future. Research in the Schools, 5, 33-38.

Thompson, B. (1990). Planned versus unplanned and orthogonal versus nonorthogonal contrasts: the neo-classical perspective. Paper presented at the annual meeting of the American Educational Research Association, Boston. (ERIC Reproduction Service No. ED318753)

Tucker, M.L. (1991). A compendium of textbook views on planned versus post hoc tests (pp. 107-118). In B. Thompson (Ed.), Advances in educational research: Substantive finding, methodological developments, 1. Greenwich, CT: JAI Press.

Table 1 - ANOVA and Follow-up Post Hoc Results for Example 1

Tests of Between-Subjects Effects

Dependent Variable: SAT94RD

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Eta Squared
Corrected Model	16414.494[a]	2	8207.247	9.542	.000	.104
Intercept	463696.553	1	463696.553	539.122	.000	.766
EXPGROUP	16414.494	2	8207.247	9.542	.000	.104
Error	141915.792	165	860.096			
Total	679816.000	168				
Corrected Total	158330.286	167				

a. R Squared = .104 (Adjusted R Squared = .093)

non-inclusion, n=53; clustered inclusion, n=31; random inclusion, n=84

Multiple Comparisons

Dependent Variable: SAT94RD

	(I) EXPGROUP	(J) EXPGROUP	Mean Difference (I-J)	Std. Error	Sig.
Tukey HSD	non-inclusion	clustered inclusion	16.47*	6.631	.035
		random inclusion	22.34*	5.145	.000
	clustered inclusion	non-inclusion	-16.47*	6.631	.035
		random inclusion	5.87	6.163	.607
	random inclusion	non-inclusion	-22.34*	5.145	.000
		clustered inclusion	-5.87	6.163	.607
Scheffe	non-inclusion	clustered inclusion	16.47*	6.631	.048
		random inclusion	22.34*	5.145	.000
	clustered inclusion	non-inclusion	-16.47*	6.631	.048
		random inclusion	5.87	6.163	.636
	random inclusion	non-inclusion	-22.34*	5.145	.000
		clustered inclusion	-5.87	6.163	.636
LSD	non-inclusion	clustered inclusion	16.47*	6.631	.014
		random inclusion	22.34*	5.145	.000
	clustered inclusion	non-inclusion	-16.47*	6.631	.014
		random inclusion	5.87	6.163	.342
	random inclusion	non-inclusion	-22.34*	5.145	.000
		clustered inclusion	-5.87	6.163	.342

Based on observed means.

Table 2 - ANOVA and Follow-up Post Hoc Tests for Example 2

Dependent Variable: SAT94LAN

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Eta Squared
Corrected Model	5094.381[a]	2	2547.190	2.460	.095	.086
Intercept	188502.386	1	188502.386	182.029	.000	.778
EXPGROUP	5094.381	2	2547.190	2.460	.095	.086
Error	53849.365	52	1035.565			
Total	275102.000	55				
Corrected Total	58943.745	54				

a. R Squared = .086 (Adjusted R Squared = .051)

non-inclusion, n=14; clustered inclusion, n=11; random inclusion, n=30

Multiple Comparisons

Dependent Variable: SAT94LAN

	(I) EXPGROUP	(J) EXPGROUP	Mean Difference (I-J)	Std. Error	Sig.
Tukey HSD	non-inclusion	clustered inclusion	23.89	12.966	.166
		random inclusion	21.27	10.416	.112
	clustered inclusion	non-inclusion	-23.89	12.966	.166
		random inclusion	-2.62	11.343	.971
	random inclusion	non-inclusion	-21.27	10.416	.112
		clustered inclusion	2.62	11.343	.971
Scheffe	non-inclusion	clustered inclusion	23.89	12.966	.193
		random inclusion	21.27	10.416	.135
	clustered inclusion	non-inclusion	-23.89	12.966	.193
		random inclusion	-2.62	11.343	.974
	random inclusion	non-inclusion	-21.27	10.416	.135
		clustered inclusion	2.62	11.343	.974
LSD	non-inclusion	clustered inclusion	23.89	12.966	.071
		random inclusion	21.27*	10.416	.046
	clustered inclusion	non-inclusion	-23.89	12.966	.071
		random inclusion	-2.62	11.343	.818
	random inclusion	non-inclusion	-21.27*	10.416	.046
		clustered inclusion	2.62	11.343	.818

Based on observed means.

Table 3

Major	examscor**	majcode1*	majcode2*	majcode3*
1	32	-1	-1	-1
1	30	-1	-1	-1
1	23	-1	-1	-1
1	27	-1	-1	-1
2	29	1	-1	-1
2	26	1	-1	-1
2	23	1	-1	-1
2	36	1	-1	-1
3	34	0	2	-1
3	27	0	2	-1
3	30	0	2	-1
3	38	0	2	-1
4	32	0	0	3
4	39	0	0	3
4	33	0	0	3
4	39	0	0	3

Columns used to "effect code" the independent variable "major".
Majcode 1 allows for groups 1 and 2 to be compared to group 3.
Majcode 3 allows for the first three groups to be compared against group 4.
**Dependent variable